Amendments to the Claims:

This listing of claims will replace all prior versions, and listings, of claims in the application:

Listing of Claims:

1. (Previously presented) A method for crawling documents, performed by one or more server devices, the method comprising:

receiving, by one or more processors associated with the one or more server devices, a uniform resource locator (URL);

receiving, by one or more processors associated with the one or more server devices, at least two different copies of a document associated with the URL; and

determining, by one or more processors associated with the one or more server devices, whether a web site corresponding to the URL uses session identifiers based on a comparison of URLs that are within the document and that change between the at least two different copies of the document, where the web site is determined to use session identifiers when a portion of the URLs that change between the at least two different copies of the document is greater than a threshold.

2. (Previously presented) The method of claim 1, wherein the document is a home page of the web site.

3. (Previously presented) The method of claim 1, further comprising:

extracting, when the URL corresponds to a web site that uses session identifiers, a session identifier from the URL to obtain a clean URL; and

determining whether the URL has already been crawled based on a comparison of the clean URL to a set of clean URLs that represent previously crawled URLs.

4. (Previously presented) The method of claim 3, wherein the compared URLs that change include URLs that are local to the web site.

5. (Previously presented) The method of claim 3, wherein the session identifiers from the URLs are extracted using rules for the web site.

6. (Previously presented) The method of claim 5, wherein the rules are determined automatically.

7. (Previously presented) The method of claim 3, further comprising:

receiving the URL as a URL from a previously crawled web document.

8. (Previously presented) The method of claim 3, further comprising:

crawling the URL when the URL is determined to not already have been crawled.

9. (Canceled)

10. (Previously presented) A method for identifying web sites that use session identifiers, performed by one or more server devices, the method comprising:

downloading, by one or more processors associated with the one or more server devices, at least two different copies of at least one document from a web site;

extracting, by one or more processors associated with the one or more server devices, uniform resource locators (URLs) from the two different copies of the web document;

comparing, by one or more processors associated with the one or more server devices, the extracted URLs of the two different copies of the document; and

determining, by one or more processors associated with the one or more server devices, whether the web site uses session identifiers when the comparison indicates that at least a portion of the URLs change between the two different copies.

11. (Canceled)

12. (Previously presented) The method of claim 10, wherein extracting URLs from the two different copies of the document includes extracting only URLs that are local to the web site.

13. (Previously presented) The method of claim 10, wherein the document is a home page of the web site.

14. (Previously presented) The method of claim 10, further comprising:

analyzing the extracted URLs, when the web site is determined to use session identifiers, to generate at least one rule identifying how the session identifiers are embedded in the URLs.

15. (Previously presented) A device comprising:

a memory to store instructions; and

a processor to execute the instructions to implement:

a spider component configured to crawl web documents associated with at least one web site; and

a session identifier component configured to determine whether the web site uses session identifiers based on a comparison of a portion of uniform resource locators (URLs) that change between different copies of at least one web document downloaded from the web site.

16. (Previously presented) The device of claim 15, wherein the spider component further comprises:

at least one fetch component configured to download content from a network; and

a content manager configured to extract URLs from the downloaded content.

17. (Previously presented) The device of claim 16, wherein the spider component further comprises:

a URL manager configured to store the extracted URLs.

18. (Previously presented) The device of claim 15, wherein the at least one web document is a home page of the web site.

19. (Previously presented) The device of claim 15, wherein the portion of the URLs that change are identified from URLs that are local to the web site.

20. (Previously presented) The device of claim 15, further comprising:

a session rule generator configured to generate rules describing how the web site embeds session identifiers in the at least one web document.

21. (Previously presented) A device comprising:

means for downloading at least two different copies of at least one web document from a web site;

means for extracting uniform resource locators (URLs) from the two different copies of the web document;

means for comparing the extracted URLs of the two different copies of the web document; and

means for determining whether the web site uses session identifiers when the comparison indicates that at least a portion of the URLs change between the two different copies.

22. (Canceled)

23. (Previously presented) The device of claim 21, wherein the means for extracting URLs from the two different copies of the web document includes means for extracting only URLs that are local to the web site.

24. (Previously presented) The device of claim 21, wherein the web document is a home page of the web site.

25. (Previously presented) The device of claim 21, further comprising:

means for analyzing the extracted URLs, when the web site is determined to use session identifiers, to generate rules describing how the session identifiers are embedded in the URLs.

26. (Previously presented) One or more memory devices containing programming instructions that, when executed by at least one processor cause the processor to perform a method for identifying web sites that use session identifiers, the one or more memory devices including:

one or more instructions to download at least two different copies of at least one document from a web site;

one or more instructions to extract uniform resource locators (URLs) from the two different copies of the document;

one or more instructions to compare the extracted URLs of the two different copies of the web document; and

one or more instructions to determine whether the web site uses session identifiers when the comparison indicates that at least a portion of the URLs change between the two different copies.

27. (Canceled)

28. (Previously presented) The one or more memory devices of claim 26, wherein the one or more instructions to extract URLs from the two different copies of the

web document includes one or more instructions to extract only URLs that are local to the web site.

29. (Previously presented) The one or more memory devices of claim 26, wherein the web document is a home page of the web site.

30. (Previously presented) The one or more memory devices of claim 26, further comprising:

one or more instructions to analyze the extracted URLs, when the web site is determined to use session identifiers, to generate at least one rule describing how the session identifiers are embedded in the URLs.

31. (Canceled)

32. (New) The method of claim 1, where determining whether the web site corresponding to the URL uses session identifiers is further based on:

determining that the at least two different copies of the document include at least two different versions of a link included in the document; and

determining that the at least two different versions of the link reference a same content.

33. (New) The method of claim 32, where determining that the at least two different versions of the link reference a same content includes:

determining that the content referenced by the link, when fetched using the at least two different versions of the link, includes different color schemes, different advertisement links, or different navigation links.

34. (New) The one or more memory devices of claim 26, where the one or more instructions to determine whether the web site uses session identifiers include:

one or more instructions to determine that the two different copies of the document include two different versions of a link included in the document; and

one or more instructions to determine that the two different versions of the link reference a same content.

35. (New) The one or more memory devices of claim 34, where the one or more instructions to determine that the two different versions of the link reference a same content include:

one or more instructions to determine that the content referenced by the link, when fetched using the two different versions of the link, includes different color schemes, different advertisement links, or different navigation links.